(12) UK Patent Application (19) GB (11) 2 376 686 (13) A

(43) Date of A Publication 24.12.2002

(21) Application No 0203140.9

(22) Date of Filing 11.02.2002

(30) Priority Data
(31) 01033646 (32) 10.02.2001 (33) GB

(71) Applicant(s)
National Institute of Agricultural Botany
(Incorporated in the United Kingdom)
Huntingdon Road, CAMBRIDGE, CB3 0LE,
United Kingdom

(72) Inventor(s)
Derek Matthews

(74) Agent and/or Address for Service
Keith W Nash & Co
90-92 Regent Street, CAMBRIDGE,
CB2 1DP, United Kingdom

(51) INT CL^7
C12N 15/10 , C07H 21/00 , C12N 15/66 , C12Q 1/68

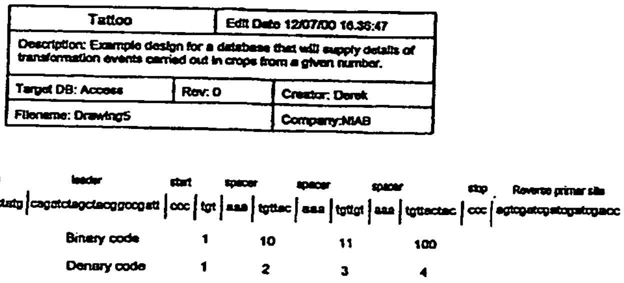
(52) UK CL (Edition T)
C3H HB4A HB7X
G1B BAC

(56) Documents Cited
WO 2001/000816 A WO 2000/068431 A
WO 2000/059917 A
J.THEOR.BIOL., VOL. 188, 1997, AJ DOIG, "IMPROVING THE EFFICIENCY OF THE GENETIC CODE BY VARYING THE CODON LENGTH- THE PERFECT GENETIC CODE", 355-60

(58) Field of Search
Other: ONLINE: EPODOC, WPI, JAPIO, BIOSIS, MEDLINE

(54) Abstract Title
Storage of encoded information within biological macromolecules

(57) Methods of encoding information within the base sequence of a nucleic acid molecule other than trough the genetic code are claimed. The embedded information may use a mechanism of Huffman coding and/or error checking, preferably by a parity check or convolutional code. The information may also be embedded in a quaternary code or by the absolute or relative lengths of restriction enzyme cleaved fragments, which preferably form a series of digits in binary or quaternary code. Also claimed are methods of decoding such information and host cells containing polynucleotides with such embedded information.



Tattoo		Edit Date 12/07/00 16.36:47
Description: Example design for a database that will supply details of transformation events carried out in crops from a given number.		
Target DB: Access	Rev: 0	Creator: Derek
Filename: Drawing5		Company:NIAB

Fig. 2 Sheet 1

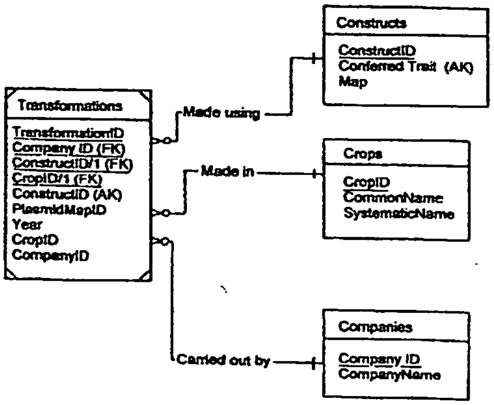


Fig. 2 Sheet 2

GB 2 376 686 A

At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

This print takes account of replacement documents submitted after the date of filing to enable the application to comply with the formal requirements of the Patents Rules 1995

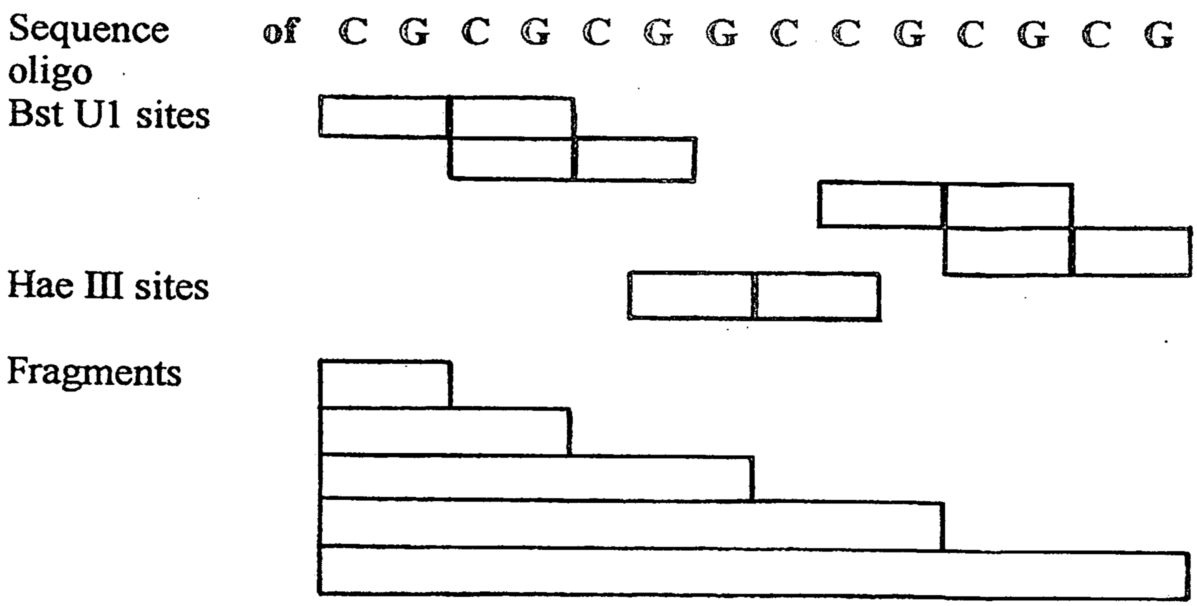

Sequence of oligo
C G C G C G G C C G C G C G
Bst U1 sites
Hae III sites
Fragments

Fig. 1

Tattoo	Edit Date 12/07/00 16.36:47	
Description: Example design for a database that will supply details of transformation events carried out in crops from a given number.		
Target DB: Access	Rev: 0	Creator: Derek
Filename: Drawing5		Company:NIAB

Forward primer site	leader	start		spacer		spacer		spacer		stop	Reverse primer site
actggtactagccagctatg	cagatctagctacggccgatt	ccc	tgt	aaa	tgttac	aaa	tgttgt	aaa	tgttactac	ccc	agtcgatcgatcgatcgacc
	Binary code		1		10		11		100		
	Denary code		1		2		3		4		

Fig. 2 Sheet 1

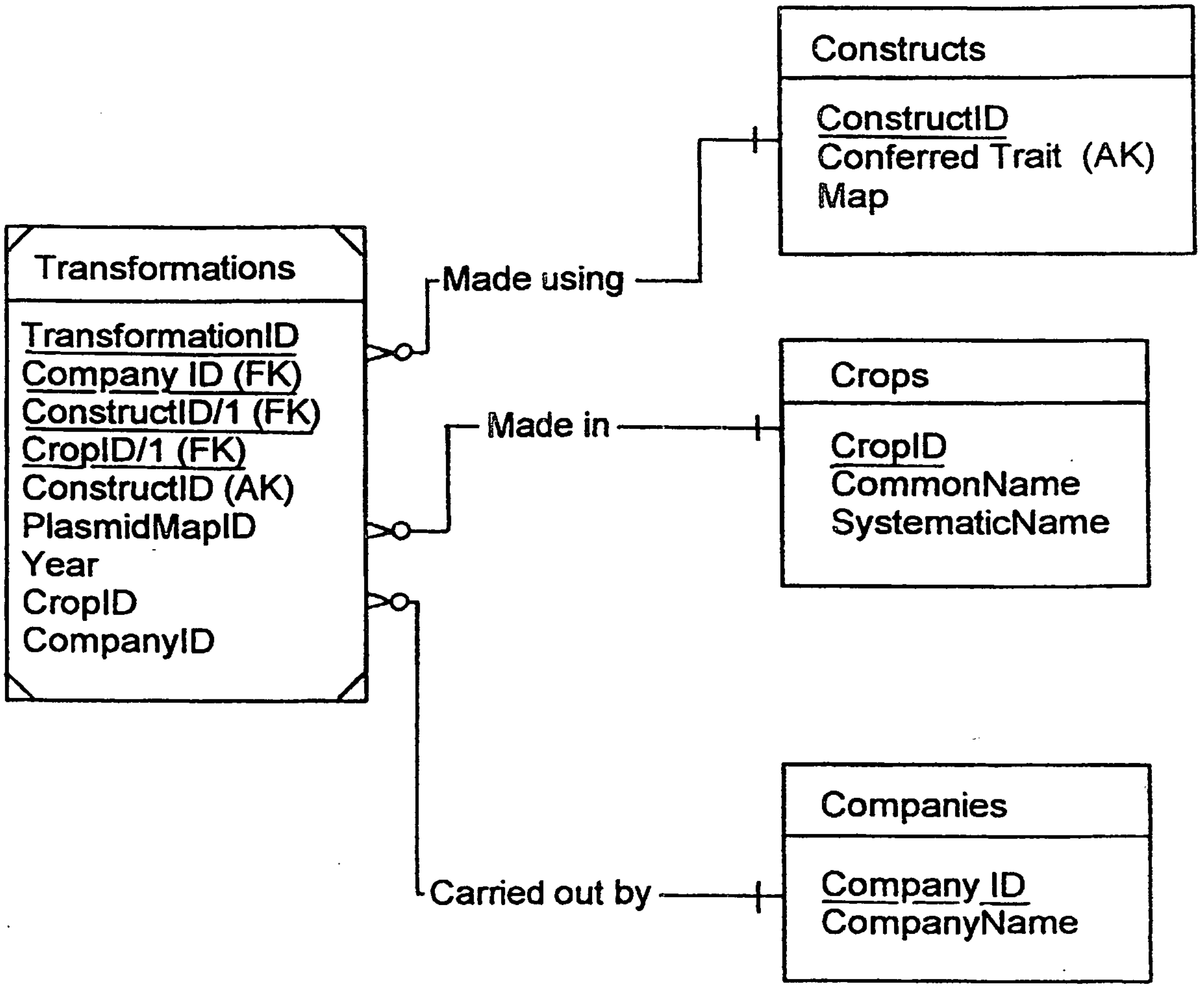

Constructs
ConstructID
Conferred Trait (AK)
Map
Transformations
TransformationID
Company ID (FK)
ConstructID/1 (FK)
CropID/1 (FK)
ConstructID (AK)
PlasmidMapID
Year
CropID
CompanyID
Made using
Made in
Crops
CropID
CommonName
SystematicName
Carried out by
Companies
Company ID
CompanyName

Fig. 2 Sheet 2

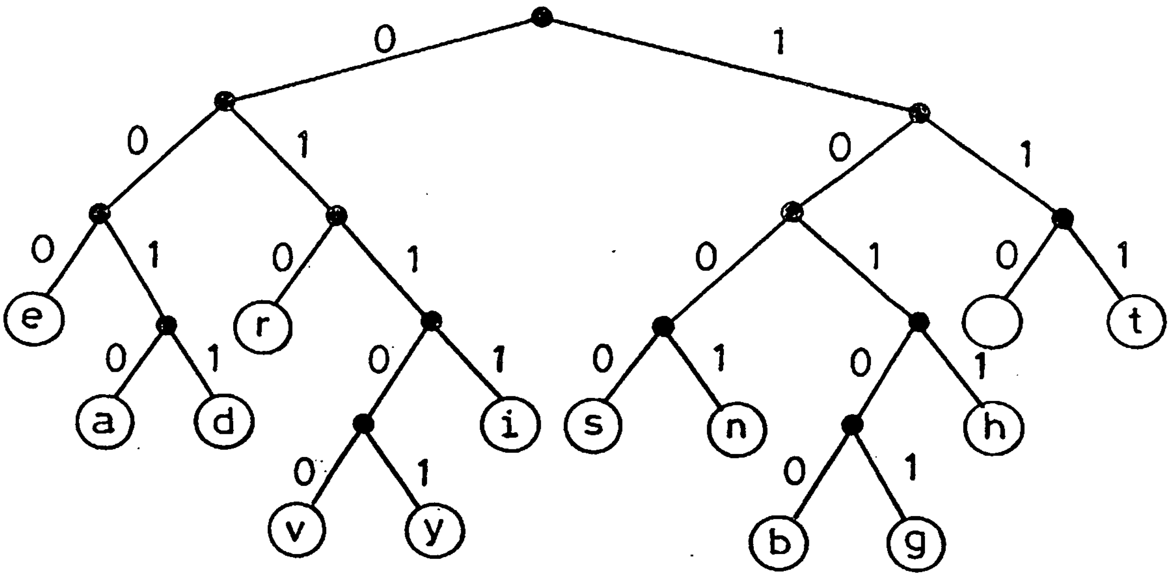


Fig. 3

Title: Storage of Encoded Information within Biological Macromolecules

Field of the Invention

This invention relates to a method of encoding information within a molecule, especially in the sequence of bases within a molecule of DNA, RNA, PNA or other polymer comprising a plurality of bases in a sequence; cells or organisms comprising a molecule carrying encrypted information; a method of decoding or decrypting information contained within a molecule; and a method of marking or tagging animate or inanimate objects.

Background of the Invention

DNA is a polymer comprising a plurality of nitrogenous bases, held in sequence on a phosphodiester backbone. In nature, DNA constitutes the genetic information store of all known organisms, with the exception of "riboviruses", which utilise RNA. RNA is essentially similar to DNA, in that it comprises a plurality of nitrogenous bases on a phosphodiester backbone.

All somatic cells in an organism contain a complete copy of all the DNA of that organism. The information within DNA is stored in the form of a sequence of nitrogenous bases; Adenine (A), Guanine (G), Thymine (T) (replaced by Uracil (U) in RNA) and Cytosine (C).

Peptide nucleic acids (PNA) are analogous to DNA/RNA in that they contain information encoded by a sequence of nitrogenous bases. They are not naturally occurring but they can be readily synthesised in the laboratory and have the advantage over DNA of being chemically more stable thus it can be used to store information in environments where DNA might rapidly degrade.

DNA and RNA may both be generally referred to as nucleic acids. Their information content is "decoded" in nature by the process of translation, by which a sequence of three bases specifies a particular amino acid to be incorporated in a polypeptide. Thus the sequence of bases in the nucleic acid determines the sequence of amino acid residues in a polypeptide encoded by the nucleic acid, the link between the base sequence and the corresponding amino acid sequence being referred to as "the genetic code", which is almost universal among living organisms.

It is known to introduce a nucleic acids into cells in order to alter the genetic constitution thereof (e.g. by introducing a gene or genes coding for a desirable polypeptide product or antisense sequences to inhibit expression of endogenous genes). It is also known to introduce "marker genes" into cells – these confer an easily detected phenotypic characteristic (e.g. antibiotic resistance), and thus serve as markers that the relevant cells have acquired the introduced nucleic acid. Techniques for performing such processes are now entirely routine for those skilled in the art.

It is also known to use DNA molecules to encode information other than by means of the genetic code.

For example, WO 01/00816 (Complete Genomics AS) discloses methods of synthesising DNA molecules which carry encoded information content. The document refers to prior art methods in which a single base encodes one or more units of information (e.g., in a binary code the base A will denote "00", C denotes "01", G denotes "10" and T denotes "11"). WO 01/00816 explains that such methods suffer the disadvantage of requiring DNA synthesis (i.e. encoding) and reading (i.e. decoding) methods in which individual bases can be discriminated. WO 01/00816 therefore emphasises that information content should be encoded by short sequences of bases rather than individual bases. The document describes in considerable detail methods of synthesising DNA molecules by assembly of pre-formed oligonucleotides. The authors do not give detailed guidance as to how to decode the assembled sequence, although they contemplate contacting the DNA molecule with a probe carrying "signalling means" (such as fluorophore). As another example (Example 7) an information-carrying DNA molecule is decoded by DNA sequence determination (i.e. a sequencing reaction).

The probe-based information retrieval taught by WO 01/00816 suffers from a number of disadvantages:
with a probe-based decoding method, each separate item of information would have to be represented by a sequence of probably at least 20 bases, in order to ensure that the sequence is unique and to ensure good specificity of hybridisation to the corresponding probe. Further, in order to ensure good specificity, it would be necessary for sequences corresponding to different units of information to differ by a reasonable amount: a single base difference, for example, between sequences 20 bases in length is unlikely to be reproducibly and reliably discriminated by probes.

Further, probe-based decoding systems are not very space-efficient, as each annealing reaction would need to be spatially-separated. A chip-based method would be space-efficient, but at present there are severe constraints on the size of probes which can be supported on chips.

WO 00/59917 similarly discloses the use of DNA molecules to encode information other than by the genetic code. The information content may be decoded by, for example, use of combinations of PCR primers to produce PCR products of different lengths (resolvable by electrophoresis), in such a way as to reveal the information content of the original DNA molecule. The DNA molecules may be used to label genetically engineered products, foodstuffs, organisms and various inanimate objects.

Summary of the Invention

In its simplest terms, the present invention is concerned with the use of macromolecules to encode information. In particular the invention is concerned with the use of a nucleic acid (which term is intended to encompass both naturally-occurring molecules, such as DNA and RNA, and synthetic molecules, such as PNA, and all variants of any of the foregoing, e.g. including non- naturally occurring bases such as modified bases, inosine, hypoxanthine and the like). Preferably the nucleic acid is a "natural" molecule (i.e. one which can be transferred to the progeny somatic cells of a genetically modified cell without a change in base sequence).

The invention relates to a method of encoding information in a macromolecule which comprises a plurality of at least 2 non-identical components which share common structural features, the sequence or order of which non-identical components carries encoded information content. Preferably the macromolecule comprises at least three non-identical components, typically four or more (such as the nucleic acid bases A, C, G and T). Preferably the macromolecule is a polynucleotide or nucleic acid, such as DNA, RNA, PNA or a hybrid of any of the foregoing. Advantageously the macromolecule is one which is capable of being replicated enzymatically e.g. by a DNA or RNA polymerase. Nucleic acids are highly preferred macromolecules as they may be stably integrated into living cells, and are amenable to manipulation and sequencing by well-known techniques. Where the macromolecule is a nucleic acid the present invention relates only to the encoding of information via a mechanism other than the genetic code.

The invention can be utilised in a great many different applications, which intended use may well affect the manner in which the invention is practised. It is appropriate therefore to describe some of the potential applications of the invention before describing in detail preferred embodiments of the invention.

In general, the invention may be used to mark or tag objects, animals, plants, products etc. with any desired information e.g. such as details of ownership, or source of origin, date of production and the like.

In general, the applications of the invention may be divided into two broad categories: those in which the macromolecule comprising the coded information is integrated into a cell or cells of an organism (preferably in a stable manner); and those in which the macromolecule is used in an essentially extracellular manner. In the former category, a particularly preferred embodiment of the invention is the use of a single macromolecule which can be incorporated into all genetically modified organisms (or, more particularly, all genetically modified plants or all genetically modified crops intended for human or animal consumption) to act as a "universal marker" of genetic modification, hence providing one easily identifiable indicator that a plant, or a foodstuff prepared from a plant, has been genetically modified.

In the latter category can be envisaged the following:

(a) use of nucleic acid 'tags' or other macromolecules to monitor the movement of people, animals or substances;

(b) use of nucleic acid labels and the like to detect pollutants and their point of origin (especially chemical or nuclear waste and the like); and

(c) use of nucleic acid or other macromolecular labels as guarantees of origin or authenticity (e.g. in foodstuffs; "designer" clothes; or as anti-fraud/security devices e.g. in bank notes, documents, credit cards).

Under (a) above are included the use of nucleic acid molecules in burglar alarms or intruder detection systems, which deposit the nucleic acid tags (e.g. in an aerosol or by other delivery system) onto unauthorised entrants into a building or room. Alternatively, dangerous or illegal goods (e.g. drugs, explosives, armaments, endangered species or products thereof) could be labelled with nucleic acid markers to track their movement. Nucleic acid labels encoding information concerning origin could provide evidence of innocence or guilt regarding theft etc.

It may be desirable to modify the nucleic acid molecule so as to inhibit or prevent digestion thereof by nucleases (especially exonucleases) or other substances which might tend to fragment the molecule, especially in those embodiments of the invention where the molecule is used extracellularly. Thus for example, modified nucleotides, especially at the 5' and/or 3' ends, may advantageously be incorporated into the molecule. alternatively, or in addition, the molecule may comprise PNA in such embodiments, as PNA is resistant to nuclease-mediated degradation.

In the other category of applications, in which the nucleic acid molecule is integrated into a cell, a variety of uses can be envisaged by the inventors. Details of ownership can be encoded in the nucleic acid and integrated into valuable cell lines (e.g. hybridomas), microorganism strains, and plasmids, cosmids or other molecules which can be introduced into cells. The nucleic acid molecules can be introduced into new transgenic organisms, especially plant varieties, again providing evidence of ownership, or providing details of the nature of the introduced trans gene. Alternatively, in many countries (especially in Europe) consumers have expressed concern about foodstuffs containing genetically modified produce. The present invention provides a simple technique of inserting a "universal" nucleic acid tag into all genetically modified plants and animals, which can be readily detected. Yet again, the invention provides a useful tool for research in ecology, especially microbial ecology and "molecular" ecology, by which specific nucleic acid marker tags can be introduced e.g. into soil bacteria or other microorganisms, or into plants, and monitoring any release of the tags into other species in the environment. The tags themselves may be very short, not coding for any polypeptide, and are therefore entirely harmless compared to conventional genetic markers, which usually confer a pronounced phenotypic characteristic.

Clearly, where it is intended that the nucleic acid molecule should be introduced into a cell there is the possibility that the molecule may be replicated and passed on in a stable manner to the progeny of the cell, without any substantial alteration in the information content of the molecule. In particular, in the case of plant cells, it may be desired for the nucleic acid molecule to become stably integrated into the cell genome. In such instances it will be preferable for the nucleic acid molecule substantially to consist of 'natural' DNA (i.e. without any non-naturally occurring bases, as these will tend to be replaced by the host cell's natural DNA repair mechanisms, thereby altering the information content of the molecule). Methods for introducing polynucleotides into cells, especially plant cells, are well known to those skilled in the art (e.g. Agrobacterium-mediated transformation, or "biolistic" techniques). Typically the polynucleotide encoding the information will be introduced as part of a vector carrying a gene or anti-sense sequence to be inserted into the cell.

In those applications in which the amount of nucleic acid molecule available for "decoding" to obtain the encoded information is small, it may be desirable for the molecule to be subjected to an amplification process. The best known of those is PCR, but several other amplification procedures are also known to those skilled in the art e.g. LCR (ligase chain reaction),"NASBA" and the like. These all require the use of one or more enzymes which, typically, will not “recognize” non-natural bases – thus, again, in such instances it will be desirable to use nucleic acid molecules substantially consisting of natural bases.

The nucleic acid molecule comprising the encoded information may desirably comprise a sequence at or near the 5’ and 3’ ends to facilitate amplification of that portion of the molecule which carries the encoded information. For example, a known primer sequence may be provided at the 5’ and 3’ ends of the molecule, (e.g. “Universal” M13 primer sequence), but in practice a unique sequence primer would be preferred.

Alternatively or additionally, the molecule may comprise a portion of known sequence which facilitates isolation of the relevant molecule from a complex mixture, quite possibly containing other, irrelevant, nucleic acid molecules. Thus, for example, the molecule may comprise a sequence complementary to an immobilised base sequence which is contacted with the mixture. The immobilised sequence may be present, for example, on a latex or magnetic particle, on an affinity chromatography column, or supported on a membrane or other solid support. Hybridisation to the immobilised sequence may therefore rapidly isolate the desired nucleic acid molecule from the mixture.

Sequences to facilitate amplification and/or isolation, as outlined above, will typically comprise at least 11 bases, preferably at least 13 bases, more preferably at least 15 bases. In the case of primer sequences they will preferably comprise at least 17 bases.

Generally the amplification/isolation facilitating sequences will not comprise encoded information, so it will be desirable to use a single universal pair of primer sequences or a single universal sequence to facilitate isolation, so that the same primers or the same ‘capture’ sequence can be used to amplify or isolate the molecule, regardless of the different information encoded by different nucleic acid molecules. It will be advantageous to use an appropriate isolation sequence which is not naturally present in the organism in question, so that the complementary capture sequence will not capture nucleic acid other than that which is associated with the encoded information.

The present inventors have devised a number of methods of encoding information into a nucleic acid molecule, and corresponding methods of decoding the molecule to reveal the information.

In a first aspect the invention provides a method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide which comprises a plurality of information encoding units, each unit being cleavable from at least one other unit by a restriction endonuclease, such that treatment of at least that part of the polynucleotide comprising the embedded information with the relevant endonuclease(s) will generate a plurality of fragments, the absolute or relative lengths of which denote the embedded information according to a predetermined relationship.

It should be noted that one pair of (typically, adjacent) information units may be cleavable by a first endonuclease, and a second pair of (e.g. adjacent) information units may be cleavable by a second endonuclease. Thus, in order to achieve a complete decoding of the polynucleotide it may be necessary, and indeed is preferable, for the polynucleotide to be treated with at least two different endonucleases. These may be contacted with the polynucleotide in a single reaction (either simultaneously or sequentially) or in separate reactions, depending on the reaction conditions under which the endonucleases exert their specific endonuclease activity. If desired more than two endonucleases may be employed.

The absolute and/or relative lengths of the fragments may conveniently be determined using PAGE. Typically in polyacrylamide gel electrophoresis (PAGE) 1 base pair resolution is obtainable up to fragment lengths of approximately 2Kb.

However, all restriction endonucleases require several bases in order to form a recognition site/cleavage site. For economy, endonucleases should preferably be selected which have very short recognition/cleavage sites. For example, *Bst*U1 cuts the sequence CG/CG, whilst *Hae*III cuts the sequence GG/CC. Other endonucleases with a 4 base recognition/cleavage site are also known.

It is particularly preferred to use a DNA molecule in which the (typically, four) base recognition/cleavage sites are overlapping, i.e. that one or more bases forming part of one recognition site also constitute part of a neighbouring recognition site. This allows for compression of more encoded information into a DNA molecule of a given length than if the neighbouring recognition sites are not overlapping. It will be noted that the recognition sites of *Bst* UI and *Hae* III have common bases, and so these can be used to provide overlapping recognition/cleavage sites.

Conveniently, the difference in length between adjacent fragments corresponds to a unit of information (e.g. a number in binary or other code), and the resulting number or number sequence can be translated by reference to the relevant database or look-up table. Accordingly, it need not be necessary to determine the absolute length of the restriction-fragments but merely their relative lengths in order to decode the encoded information.

The polynucleotide may be prepared by any suitable method, such as conventional *in vitro* synthesis, or by assembly of sub-components by ligation, by PCR etc. or by means of the method disclosed in WO 01/00816.

In a second aspect the invention provides a method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide, subjecting at least that part of the polynucleotide carrying the embedded information to digestion by at least one (preferably two, or more) restriction endonucleases, so as to generate a plurality of restriction fragments, and determining the absolute or relative lengths of the fragments, wherein the lengths of the fragments, or the differences therebetween, have a predetermined meaning.

In a third aspect, the invention provides a method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide wherein the embedded information is encoded (and decodable) by a mechanism which includes the use of Huffman coding and/or error checking.

In a fourth aspect the invention provides a method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide, and decoding the embedded information by a mechanism that involves the use of Huffman coding and/or error checking.

In the third and fourth aspect of the invention the information may be embedded by any means e.g. by a letter or number substitution system, by a PCR or restriction fragment length system or, for example, by any of the methods disclosed in WO 01/00816 or WO 00/59917.

Huffman coding is a system which is well known to those skilled in the art of computer science. However, no one has hitherto considered applying Huffman coding to information embedded in polynucleotides, as the field of molecular biology is traditionally remote from that of computer science. Huffman coding is explained in detail in Binstock & Rex (1995 pp 490-500 "Practical Algorithms for Programmers [Reading, M A: Addison-Wesley]) and Nelson & Jean-Loup (1996 pp 31-34 "The Data Compression Book" [New York: M & T Books]). In essence, Huffman coding is a data compression technique by which one can produce size-efficient codes for data for which one has statistical information regarding the frequency of occurrence of the symbols, so that the most frequently occurring symbols (e.g. the letter 'e') are represented by the shortest motif, whilst rarely occurring symbols (e.g. the letters 'v' and 'z') are represented by larger motifs.

In computer science, Huffman coding has traditionally been applied as a data compression technique. In the context of the present invention, Huffman coding has the additional advantage that it allows information to be encoded in the polynucleotide by "words" (i.e. sequences) of variable length.

Similarly error checking is a general principle which is well known to those in the field of computer science but is not a technique which is normally associated with molecular biology. Particularly preferred error checking techniques in the context of the present invention are parity checks and convolutional codes. These two techniques are described in greater detail in the examples.

Huffman coding and error checking techniques such as parity checks and convolutional codes are especially (but not only) useful in embodiments in which the embedded information is retrieved by sequence determination of the polynucleotide.

Methods of sequence determination are of course well-known and routine for those skilled in the art. In such embodiments it is advantageous if the portion of the polynucleotide requiring sequencing can be kept below about 2000 bases, (preferably less than about 500 bases) as this is the upper limit of the size of sequence which can be determined on a single sequencing run. Obviously, for longer messages, overlapping sequencing can be performed, using "primer walking" (in which sequencing primers are prepared using knowledge of the sequence at the end of a preceding sequencing run to extend the sequence in a subsequent sequencing run).

In certain embodiments, the determined base sequence may not directly correspond to the sequence of letters in the decoded message. For example, the invention may employ a number substitution system in which particular base sequence motifs correspond to a predetermined number, typically in a mathematical system other than base 10 (i.e. a non-decimal system). Binary, ternary or, especially, quaternary-based codes will generally be preferred. The resulting number, or number sequence, may then be translated by reference to a database or look-up table.

It should be appreciated that the amount of space in the nucleic acid molecule available for accepting embedded information may well be limited. For example, where the information is encoded in a vector used to introduce a desired gene or anti-sense sequence into a host cell, much of the space in the vector may be taken up by other essential features, leaving comparatively little space for accepting embedded information. In any event, the smaller the sequence to be determined, the faster and easier the information can be decoded.

Methods of the invention in which the encoded information is decoded by sequencing, in particular, lend themselves to the application of other data handling and data processing techniques practised in other fields of technology, such as error correction techniques like parity checks and convolutional codes. Such techniques may thus be employed in the method of the invention, as described in the Examples below.

In a fifth aspect the invention provides a method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide wherein the embedded information corresponds to a sequence of digits in a quaternary code, each digit or combination of digits having a predetermined meaning.

In a sixth aspect the invention provides a method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide and treating it in such a way as to retrieve the embedded information as a sequence of digits in a quaternary code, each digit or combination of digits having a predetermined meaning.

The inventors have found that quaternary code lends itself particularly to embedding information in polynucleotides for incorporation in living organisms, since naturally occurring DNA molecules comprise four different bases.

In the illustrative Examples which follow, the invention is illustrated in the context of providing information regarding transgenic plant varieties, in particular providing information regarding (i) the name of the company or organisation responsible, (ii) the plant species, (iii) the identity of the nucleic acid construct introduced into the plant and (iv) the year the variety was produced.

The invention will also be described by way of illustrative drawings. In the accompanying drawings:

Figure 1 is a schematic representation of retrieval of embedded information from a DNA molecule by digestion of the molecule with a pair of restriction endonucleases to generate a plurality of restriction fragments, which fragments are then resolved (e.g. by PAGE) and the relative size of the fragments determined, which corresponds to a series of digits (typically in binary or quaternary code);

Figure 2 is a schematic representation of the sort of information which may be embedded in a polynucleotide construct inserted into a genetically modified plant; and Figure 3 is a schematic diagram of a "tree" used in encrypting and decrypting a message by a mechanism involving Huffman coding.

Example 1

Substituted Alphabets

In a substitution alphabet coding approach, DNA sequence motifs are used to represent each letter of the alphabet. As there are four bases the smallest motif size that will allow substitution of all the letters of the English alphabet is a triplet motif. The DNA is sequenced and the code read back directly. The system would contain two components: the sequence that carries the actual code and a pair of PCR primer sites that allow the DNA to be extracted from the genome and also allow the code to be sequenced.

Such an approach has the advantage of conceptual simplicity, but would be very inefficient for most messages, and there is the possibility of secondary structures occurring in the molecule (depending on the encoded message), which could block sequencing reactions.

Many of these problems can be overcome, or at least ameliorated, by the use of Huffman coding, so that information can be embedded in the smallest possible length of sequence.

Example 2

Restriction Fragment length encoding

This technique relies on encoding information in the form of restriction fragment lengths or fragment length differences.

Sequences performing this function would contain three functionally distinct regions:
(i) primer sites (typically 15 bp); (ii) restriction sites (typically 4-6 bp); and (iii) an optional filler sequence, depending on the resolution of the gel system to be used.

Fragment length differences are used to denote the coding elements (e.g. 0 or 1). Practical considerations of resolution determine the size difference used and the maximum amount of information that can be carried.

Using PAGE, 1 bp resolution is available up to approximately 2 Kb and therefore code elements can comprise increments of 1 bp (e.g. 1 bp = 0 and 2 bp = 1). Restriction sites however add a certain overhead. For example *Bst* U1 cuts the following sequence CG|CG whilst *Hae* III cuts the sequence GG|CC. It is possible using these two recognition sequences to build an overlapping code that gives fragment differences of 2 and 3 bp. The arrangement is illustrated schematically in Figure 1.

Using this approach each fragment to be read in turn must exceed the length of the preceding fragment by the size of the symbol it has to represent. For a string of zeros in a binary system, for instance, the first fragment would be 2 bp long, the second 4 bp and so on. A message of 10 information units therefore requires 20 bp of sequence. The longest message that can be carried by this system is around 500 information units in length.

The advantages of such a method include obviating the need for sequencing reactions, hence rapid analysis is possible. There should be low error rates, and the detection/analysis methods are cheap to perform.

However, an amplification step will often be required, in order for there to be sufficiently large amounts of DNA to digest and to give rise to detectable amounts of restriction fragments.

Example 3

Number system substitution

This is a method for encoding numeric data into an oligonucleotide in such a way that it can be easily recovered and read from a complex genetic background or other substrates. By way of example, a binary system will be considered first.

General description

A DNA sequence of between 39 and 1000 bp (or whatever the upper sequencing limit is) is provided comprising a number of distinct functional units. At each end of the DNA sequence are regions that enable the DNA to be amplified, by PCR, from a complex genetic background or other substrate, such as processed food. These terminal regions also serve as primer sites for sequencing the intervening DNA. Proximal to the terminal primer sites are a start and stop signal respectively comprising a DNA triplet, which define the information containing region. Between the start and stop signals are one or more (depending on the application) reading frames that contain a binary number encoded in the form of a DNA triplet. A triplet substitution is preferably chosen to avoid microsatellite-like dinucleotide repeats. If there is more than one reading frame, between each reading frame there is a triplet spacer that indicates the extent of the reading frame and prevents errors. In order to recover this molecule and read the data encoded therein the following would be carried out:

i) DNA isolation from the organism / product /subject of interest;

ii) PCR amplification of the whole construct;

iii) Bi-directional di-deoxy sequencing of the construct;

iv) Conversion of the triplet code sequence into binary digits; and

v) Comparison with an appropriate database or lookup table to decode the information held within the construct.

Description of components

Left and right primers and leader sequences

The left and right primer sequences are 20 bp regions of unique and known sequence. During PCR amplification, PCR primers anneal to these regions to direct exponential amplifiction of the whole DNA sequence. By ensuring that the primer sites have a different sequence, an orientation is given to the whole molecule. Proximal to the primer annealing sites are a short region (10-20 bp) of arbitrary leader sequence that facilitates the later sequencing of the whole molecule and ensures that the whole of the information carrying portion is sequenced.

Start and stop signals

These triplets unambiguously mark the beginning and end of the information-containing region especially where the sequence at the beginning of the read is of poor quality or truncated. The triplet CCC is chosen as it doesn't occur at any junction of the triplets encoding the binary numbers and is unlikely to arise due to reading errors, frame shifts or mutation of the triplets that comprise the binary code.

Spacers

These triplets delineate the borders of each reading frame and allow the binary word length in the reading frames to be variable. The triplet AAA is chosen as it doesn't occur at any junction of the triplets encoding the binary numbers and is unlikely to arise due to reading errors, frame shifts or mutation of the triplets that comprise the binary code.

Reading frames

These comprise a region of variable length made of a string of the DNA triplets TGT and/or TAC. If the triplet TAC is made to stand for the binary digit 0 and TGT for the digit 1, a binary (base 2) number can be read from the string of triplets. Triplets are chosen to avoid the possibility of tandem dinucleotide repeats that would cause problems during PCR (stuttering) and sequencing. These particular triplets are chosen as they are nonsense codons within the genetic code (i.e. do not encode a particular amino acid). This reduces that likelihood of extensive homology with regions of the plant's natural genome. The use of these triplets also reduces the potential for secondary structure formation.

A theoretical example of such a system is shown in Figure 2.

Binary systems of this sort may be useful. However considerable space can be saved by the use of higher order number systems. The four base composition of natural DNA appears to the inventor to lend itself to adopting a quaternary number substitution system. The following table illustrates the space requirements of the first three number systems as compared to decimal:

Decimal	Binary	Ternary	Quaternary
0	0	0	0
1	1	1	1
2	10	2	2
3	11	10	3
4	100	11	10
5	101	12	11
6	110	20	12
7	111	21	13
8	1000	22	20
9	1001	100	21
10	1010	101	22

From this it can be seen that the use of a quaternary number system in the above example would offer a 50% space saving over the use of a binary substitution. In addition as this is not a binary system it is less repetitive and does not require the use of a triplet code; for example the following could be used.

Quaternary code	A	T	C	G
Decimal number	0	1	2	3
0				A
1				T
2				C
3				G
4			T	A
5			T	T
6			T	C
7			T	G
8			C	A
9			C	T
10			C	C
11			C	G
12			G	A
13			G	T
14			G	C
15			G	G
16		T	A	A

Example 4

Coding for error correction

Parity checks

Many techniques for coding to reduce errors are based on the concept of parity checks. For example, the parity of a set of binary digits may be defined as 1 if the set contains an odd number of 1s and defined as 0 otherwise. Thus, the parity of 111 is 1, that of 110 is 0, that of 0001 is 1, etc. A change of any single digit (from 0 to 1 or *vice versa*) in the set invariably changes the parity of the set. Thus, if a set of digits is transmitted, followed by transmission of the parity of the set (called a parity-check digit), a single error in the transmission can be detected by the disagreement between the parity of the received set and the received parity-check digit. Of course, this disagreement does not indicate which received digit is in error, and no correction is possible.

The following example shows how parity checks can be used to correct errors. Four source digits (s_1, s_2, s_3, s_4) are transmitted, followed by three parity-check digits, p_1, p_2, p_3. The digit p_1 is chosen as the parity of the first three source digits; p_2 as the parity of the first, second, and fourth; and p_3 as the parity of the first, third, and fourth. Then, for example, the source digits 1001 would be encoded into 1001100. If the first source digit s_1 is received in error at the receiver but all other digits are received correctly, then each received parity check will disagree with the parity of the corresponding set of received source digits. Similarly, if the second source digit is received in error, only the first two parity checks will disagree with the corresponding received parities, because s_2 is not in the set used for the third parity check. Similarly, each source digit and parity-check digit has its own unique pattern of disagreements, allowing any single error in the seven digits to be corrected. Multiple errors unfortunately cannot be corrected in this example.

This type of code is an example of a block code, a code in which the source digits are segmented into sequences of a fixed length (4, in this case), and there is a rule for transforming each source sequence into a sequence of digits of fixed length (7, in this case). The block length of the code is defined as the length of the sequence, and the code rate (assuming a binary source) is the ratio of source sequence length to final sequence length. Thus, the code rate for the example is 4/7.

Convolutional codes

Many practical applications of coding to correct errors use convolutional codes instead of block codes. For example, suppose that each source digit being encoded is followed by a parity-check digit so that the final sequence has the form $s_1\, p_1\, s_2\, p_2\, s_3\, p_3 \ldots$ Suppose that each parity-check digit is the parity of the two preceding source digits. Thus, if the first four source digits are 1011, the first eight transmitted digits will be *1*1*0*1*1*1*1*0* (the source digits are italicised).

If a sequencing error alters one of the source digits, then the parities will not check for the two succeeding parity-check digits. On the other hand, if sequencing alters a parity-check digit, only that one parity will fail to check. Thus, any single error (if preceded and followed by at least three correct digits) can be identified and corrected.

In order to attain very low probabilities of decoding error with convolutional codes, the parity-check constraints must extend over more digits. Rather than making p_n the parity of s_n and s_{n-1}, p_n should preferably be the parity of a large set of previous source digits in a fixed set of positions relative to p_n.

Both parity checks and convolutional codes could be applied as error traps for codes within DNA.

Example 5

Detection of GM material.

Currently in Europe there is segregation between GM and non-GM produce. There is a need for food products and seeds to be screened to demonstrate 'freedom from GM contamination'. Currently this is achieved by searching for DNA containing commonly used promoter and terminator sequences. At present this approach is workable because the number of promoters and terminators in general release is small. Looking to the future there are a wide range of new promoters and terminators that will be used, this will make effective screening progressively more difficult.

If all constructs inserted into plant varieties contained a common DNA sequence (e.g. the left and right primer sites proposed in Example 3) then there would be a simple and universally applicable method for detecting GM material; amplification from the universally used primer sequences would denote the presence of a genetic modification.

This argument can also be applied to monitoring the 'escape' of genetically modified organisms or constructs into the natural environment.

Identification of the GM Trait or Construct.

Following amplification of the GM construct sequencing of DNA would allow the coded information to be read. In example 3 the code would be in the form of four or more separate numbers which would identify the construct unambiguously. For example:

Base Sequence	Binary Code	Database field (decimal equivalent)
Spacer		
TGT	1	Company name (1)
Spacer		
TGTTGT	11	Species (3)
Spacer		
TGTTGTTGT		
TGTTGTTAC		
TGTTACTAC		
TACTGT	11111010001	Year (2001)
Spacer		
TGT	1	Construct (1)

In this example the base sequence shows that the construct; comes from the company coded '1', is inserted into the species coded '3', was registered in the year 2001 and is coded construct '1' from that company in that year. A central registration database might reveal this to be:

Company:	NIAB
Species:	Maize
Year:	2001
Construct:	1 (detail of the composition of the construct and its action would be held in confidence).

Thus the presence of this construct in a processed food would show that GM maize with this genetic manipulation was present. If the construct were not approved for release it would show infringement of regulations.

In seed the presence of the construct in all seeds in an appropriate sample would show that the seed was genetically modified and that it contained a specific trait. This would give assurance to buyers and end users that they were obtaining seed with the required characteristics.

Example 6

Standard Molecular Biological Procedures

6.1 A general purpose method for the Extraction of DNA from plant material.

DNA may be extracted from a variety of plant tissues using this procedure modified from Murray and Thompson (1980 Nucl. Acids Res. 8, 4321-4325). Fresh or frozen leaf material is ground with liquid nitrogen, in a mortar and pestle, to a fine light green powder. The powder is then transferred to a 50 mL centrifuge tube. Preheated (65°C) extraction buffer [1.4 M NaCl, 1% (w/v) CTAB, 50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 1% dithiothreitol (DTT)] is then added to a total volume of 20 mL and incubated for 30 min at 65°C with occasional shaking. Cell debris is removed by centrifuging at 3500 x g for 10 min and the supernatant extracted twice with an equal volume of chloroform:isoamyl alcohol (24:1). The supernatant is then mixed thoroughly and the phases separated by centrifugation for 5 min at 3500 x g. The aqueous (upper) phase is retained and 0.1 volume of 10% (w/v) CTAB, 0.7 M NaCl added. This is then mixed thoroughly, and extracted with an equal volume of chloroform:isoamyl alcohol (24:1) as above. One volume of precipitation buffer [1% (w/v) CTAB, 50 mM Tris-HCl (pH 8.0) and 10 mM EDTA] should then be added. The DNA-CTAB complex is precipitated at 15°C for 30 min, pelleted by centrifugation at 3500 x g for 10 min and redissolved in 500 µL high salt buffer [10 mM Tris-HCl (pH 8.0), 5 mM EDTA and 1 M NaCl]. Particulate material is removed by centrifugation at 13000 x g for 5 min, the DNA is then precipitated with 0.8 volumes of isopropanol and centrifuged at 13000 xg for 15 min. The DNA pellet is then washed once with 70% (v/v) ethanol and once with absolute ethanol, briefly air dried and then dissolved in 100 µL TE [10 mM Tris-HCl (pH 8.0), 1 mM EDTA]. Heat treated RNase A is then added to a final concentration of 10 µg/mL and incubated at 37°C for at least 30 min. The reaction is stopped by extracting once with chloroform:isoamyl alcohol (24:1) and the aqueous phase recovered by centrifuging at 13000 xg for 5 min. The aqueous layer is transferred to a fresh tube and 0.1 volume of 3.0 M sodium acetate (pH 5.2) and 2.5 volumes of absolute ethanol added and mixed well. The DNA is then sedimented by centrifuging at 13000 x g for 15 min, air dried for 30 min and redissolved in 50 μL TE buffer. DNA concentration and purity may then be assessed as described below.

6.2. DNA Quantitation and assessment of Purity

DNA concentration may be measured by ultraviolet absorbance spectrophotometry. The concentration of DNA in a sample is directly proportional to the amount of ultraviolet radiation absorbed by the solution. At a wavelength of 260 nm, an absorbance of 1.0 corresponds to 50 μg of double-stranded DNA per mL of sample. In addition to measuring DNA concentration, ultraviolet absorbance can be used to check the purity of a DNA sample. The ratio of absorbance at 260 nm compared with 280 nm can give an indication of the purity of the sample. A ratio of 1.8 indicates that the DNA preparation is pure. A ratio of less than 1.8 indicates that the preparation is contaminated with protein or phenol, and a ratio approaching 2 indicates RNA contamination (Sambrook *et al,* 1989). A useful measure of DNA purity can be attained by measuring the absorbance between the ranges of 220 nm and 320 nm. By plotting the absorbance against the wavelength, the curve gained can denote the level of purity of the DNA sample. Ten microlitres of the DNA prep are diluted in 1 mL TE buffer (dilution factor of 100) in a 1 mL quartz cuvette (Shimadzu). Absorbance measurements are taken at wavelengths of 260nm and 280 nm and compared with the readings taken of the pure TE buffer (to zero the instrument). DNA concentration is estimated using the following calculation:

$$A_{260} \times 50 \times dilution\ factor = DNA\ concentraction(\mu g\ /\ mL)$$

$$A_{260}\ x\ 50\ x\ dilution\ factor\ =\ DNA\ concentration\ \ (g/mL)$$

6.3. A general procedure for restriction digestion of DNA

The following is a general method for restricting DNA with the enzymes *Bst* U1 which cuts the sequence CG|CG and *Hae* III which cuts the sequence GG|CC. Ten micrograms of DNA are placed in a microfuge tube and water added to a total volume of 18 µL. Two microlitres of 10 X restriction buffer (supplied by the manufacturer) and 1-2 units of restriction enzyme are added and mixed. The samples are then incubated at 37°C for at least 1 hour, after which the reactions are stopped by adding 0.02 volumes 0.5 M EDTA (pH 8.0). The enzyme is then removed by extracting with an equal volume of chloroform:isoamyl alcohol (24:1). The DNA is then precipitated by adding 0.1 volumes of 3.0 M sodium acetate (pH 5.2) and 2.0 volumes of ice-cold ethanol. The tubes are stored on ice for 1 h, and the DNA pelleted by centrifugation at 13000 x g for 15 min. The pellet is then washed with 70% (v/v) ethanol and re-sedimented by centrifugation at 13000 x g for 2 min. The DNA pellet is air dried for 30 min and then redissolved in 10 mM TE [0.01 M Tris-HCl (pH 8), 0.001 M EDTA] buffer at a final concentration of 100 µg/mL. DNA fragments may be separated using agarose gel electrophoresis (see 7.9), and blotted onto nitrocellulose membranes using the technique of Southern (1975 J. Mol. Biol. 98, 503-507), (see 6.4 below).

6.4. Immobilisation of DNA fragments on to membranes (Southern Blotting)

DNA fragments separated in polyacrylamide or agarose gels may be transferred (blotted) onto nitrocellulose or nylon membranes using the procedure described by Southern (1975). An agarose gel is covered with depurination solution (0.25 M HCl) and agitated gently until a point 15 min after the dye has changed colour from blue to yellow. The depurination solution is then removed and the gel rinsed briefly in distilled water. The gel is then covered with denaturation solution (0.5 M NaOH, 1.5 M NaCl) and agitated gently for 20-25 min, before being rinsed briefly in distilled water. The gel is then covered with neutralisation solution [0.5 M Tris-HCl (pH 7.5), 1.5 M NaCl] and agitated gently for 15 min. The neutralisation solution is then replaced with fresh neutralisation solution for a further 15 min. The capillary blot apparatus should then be assembled, and left overnight for the DNA to transfer onto the membrane (e.g. Tropix Tropilon Plus, PE ABI). The membrane is then washed in 2 X SSC for 5 min to remove any agarose, and the DNA is then covalently bound to the membrane by baking at 80°C for 2 hours and then cross-linking on an ultraviolet transilluminator (302 nm, Ultra Violet Products model GDS 7500) for 5 min.

6.5. Dot Blots

DNA samples are diluted to approximately 2 µg/µL. One microlitre aliquots are pipetted onto a membrane (e.g. Tropix Tropilon plus) and allowed to dry. The membrane is then covered with denaturing solution (as described above) for 5 min followed by neutralisation solution (as above) for 5 min. The DNA is then fixed to the membrane by ultraviolet crosslinking as described in Section 6.4.

6.6. Hybridisation of Probes

Dot blots and Southern blots may be probed with biotin labelled oligonucleotides using the following method. Membranes are placed in gauze bags (400 M mesh size, Hybaid, UK), inside polyethene bags. 10 - 15 mL of preheated hybridisation buffer {10% (w/v) PEG 600, 7.5% (v/v) 20 X SSPE [3 M NaCl, 0.2 M $NaH_2PO_4.H_2O$, 0.02 M EDTA, (pH 7.4)], 7% (w/v) SDS} is added to the bag, which is then heat sealed (HM2500 polythene heat sealer, Jencons Scientific Ltd) and prehybridised for 1 hour at 55°C. The membranes are then hybridised overnight at 55°C in a water bath with 250 µL probe mixture (0.1 - 1 pMol/ mL (hybridisation buffer)). They are then washed twice in 2 X SSC, 0.1% (w/v) SDS at room temperature for 5 min, followed by two 5 min 45°C washes in 1 X SSC, 0.1% (w/v) SDS. The presence of bound probe may be detected using a chemiluminescence procedure, as described below.

6.7. Chemiluminescence Detection

To detect biotin labelled oligonucleotides a chemiluminescence procedure may be used (Matthews *et al.*, 1997). Following hybridisation the still wet membrane is washed for 20 min in blocking buffer [0.1% (v/v) Tween 20, 0.2% (w/v) I-block reagent (Tropix), 58 mM Na_2HPO_4, 17 mM NaH_2PO_4, 68 mM NaCl]. Avidin-alkaline phosphatase conjugate [1 mg/ mL in 50 mM Tris-HCl (pH 7.5), 0.2 mg/ mL sodium azide] is diluted 1: 20 000 in 80 mL conjugate buffer [0.2% (w/v) I-block reagent, 58 mM Na_2HPO_4, 17 mM NaH_2PO_4, 68 mM NaCl]. The blocking buffer is poured off, replaced with conjugate buffer and incubated at room temperature for 20 min. The membrane is then washed once in blocking buffer for 5 min and three times in wash buffer (0.3% (v/v) Tween 20, 58 mM Na_2HPO_4, 17 mM NaH_2PO_4, 68 mM NaCl) for 5 min. The membrane is then equilibrated to pH 9.5 by two five minute washes in assay buffer [(50 mM sodium carbonate, 1 mM $MgCl_2$) added to (50 mM sodium bicarbonate, 1 mM $MgCl_2$) to give pH 9.5] and transferred to a solution comprising 25 mM CDP-*Star*®substrate [disodium2-chloro-5-(4-methoxyspiro{1,2-dioxetane-3,2'-(5'-chloro)-tricyclo[3.3.1.$1^{3.7}$]decan}-4-yl)-1-phenyl phosphate], (PE ABI) for 5 min. The wet membrane is then sealed between two OHP sheets (Dudley write-on OHP film) and incubated for 1 hour in the dark. The membrane is then placed between two sheets of Fuji Rx X-ray film (GRI) for approximately 5 min. The film is then removed and placed in developer (Photosol CD18) for 5 min. The film should then be rinsed in water and transferred to photographic fixer (Photosol CF40) for several minutes (until the film has become clear). The film is then rinsed in water and left to dry.

6.8. Procedure for carrying out PCR reactions

PCRs are performed in 25 µL volumes comprising 1 unit Taq polymerase (BioGene), 1mM Tris-HCl (pH8.3), 5 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM each dNTP (Gibco Life Technologies), 1 µM primer(s) (Genosys Biotechnologies or MWG Biotech), and 30 ng of target DNA.

Reactions may be carried out either in 0.5 mL microtubes in a MJ Research PTC-100, or in 0.2 mL tubes in a Techne Genius or a PE ABI 9700. A typical PCR programme may comprise an initial denaturation step at 95°C for 5 min, followed by 30 cycles of [30 sec at 95°C, 30 sec at a specific annealing temperature, 1 min at 72°C] followed by a final 10 minute extension at 72°C. The annealing temperature will vary according to the primer(s) used in the reaction.

6.9. A typical Procedure for carrying out Agarose Gel Electrophoresis

Agarose gels should be run in 0.5 X TBE buffer [45 mM Tris-borate (pH 8.0), 1 mM EDTA]. To make a 1% mini agarose gel, 0.5 g Seakem LE agarose (FMC Bioproducts) is added to 50 mL of 0.5 X TBE buffer and heated for 2 to 3 min, with frequent, gentle agitation, at full power in a 650 W microwave until the solution is boiling and all the agarose is completely dissolved. Masking tape is used to seal either end of a 7 x 10 cm submarine horizontal gel tray (Flowgen) before the cooled (approximately 50°C) gel mixture is slowly poured and allowed to set. While the mixture is still molten, a 1 mm 12 sample well comb is positioned in the comb slots, 1 cm from one end of the gel tray. Once the gel has completely set, the tape and comb are carefully removed and the tray placed in the horizontal gel electrophoresis unit (Flowgen) and completely covered in 0.5 X TBE buffer to a depth of 0.5 cm.

DNA samples are mixed with a loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol, 15% Ficoll) in a 5 to 1 ratio. The samples are then loaded through the TBE buffer into the wells in the gel. To determine the size of the sample, marker DNA (e.g. 100 bp ladder, Amersham Pharmacia Biotech) is also loaded into a separate well. A voltage of approximately 5 V/ cm is applied across the gel using a BioRad power pack 300. The DNA is allowed to migrate until the bromophenol blue dye is approximately 1 cm from the end of the gel (1 to 2 hours).

In order to visualise the DNA in the gel, the gel should be removed from the tank and placed in a 0.5 µg/ mL solution of ethidium bromide for 20 min. The gel is then destained in distilled water for a further 10 min and viewed on an ultraviolet transilluminator (e.g. 302 nm, Ultra Violet Products model GDS 7500).

6.10. Polyacrylamide Gels

Polyacrylamide gels may be prepared from ready to use sequencing gel solutions, Sequagel®-6 (National Diagnostics), according to the manufacturer's instructions: gels are prepared by mixing 40 ml sequagel monomer concentrate with 10 ml sequagel complete buffer, 400 µl freshly made 10% (w/v) ammonium persulphate is added to initiate polymerisation. This mixture is used to cast the gel between 2 clean glass plates and left to polymerise for 1 to 2 hours. Short acrylamide gels (20 cm) are cast using 1 mm spacers and run in a Protean®-II xi cell (BioRad), for 2 hours at 18 mA constant current, in 1 x TBE buffer. Large sequencing gels (40 cm) are cast using 0.4 mm spacers and run in a SQ3 (Pharmacia Biotech) gel tank for 2 hours at 25 volts/ cm in 1 X TBE buffer. Gels to be silver stained, are immediately placed in 10% (v/v) glacial acetic acid. Gels to be analysed using chemiluminescence are removed from the glass plates, by gently pressing dry Whatmann 3 MM chromatography paper onto the surface of the gel, and peeling both the paper and the gel off the glass plate. Tropix Tropilon membrane (soaked in 20 X SSC) is placed in contact with the top of the gel and a glass plate placed on top to ensure even contact between the gel and the membrane. The PCR products are allowed to transfer to the membrane for 1 hour, before being fixed in place by cross-linking on an UV transilluminator (302 nm, Ultra Violet Products model GDS 7500) for 5 min. The chemiluminescent procedure is described in Section 6.7.

6.11. Silver Staining

The procedure for silver staining is taken from Bassam *et al.* (1991 Anal. Biochem. <u>196</u>, 80-83). The gel is fixed in 10% (v/v) glacial acetic acid for 20 min or until tracking dyes are no longer visible, whichever is longer. The gel is washed 3 times in distilled water for 2 min each before being transferred to the staining solution [6 mM $AgNO_3$, 0.05% (v/v) formaldehyde] for 30 min. The gel is washed briefly in water and the developer [0.3 M Na_2CO_3, 0.05% (v/v) formaldehyde, 2 mg/ml sodium thiosulphate] added immediately. The gel is then transferred to fresh developer once the bands start to appear, for 2 to 3 minutes while the band intensity increases, and then subsequently fixed in 10% (v/v) glacial acetic acid for 5 min. To preserve the gel, it should be placed in 8.7% (v/v) glycerol for 1 hour and then covered with pre-cut cellophane sheets, (Pharmacia Biotech) and left overnight at 37°C.

6.12. LI-COR Gels

Samples are denatured at 95°C for at least 3 min and 0.3 µL loaded onto an 18 cm sequencing gel. The sequencing gel comprises 6 ml Long Ranger (Flowgen), 21 g urea, 6 mL 10 X long run TBE (1340 mM Tris base, 450 mM boric acid, 25 mM EDTA) and 22 ml H_2O. The solution is degassed for 15 min before adding 33 µl TEMED and 333 µl freshly made 10% (w/v) ammonium persulphate. Gels are left to set for at least 2 h before electrophoresis. The default electrophoresis conditions for 18 cm gels using Base Image IR data collection software are 1500 V, 35 mA, 40 W, 48°C, 25 frames, scan speed 3, prerun until plate temperature was 48°C, in 1 X long run TBE.

6.13. Sequencing of Inserts

DNA sequences may be characterised using ABI 377 automated sequencers and according to the protocols supplied by PE ABI, or using the LI-COR Gene ReadIR^2 DNA Analyser system. Reaction master mixes comprise template DNA (approximately 750 ng) with 1.4 µL DMSO, 1.5 pmol M13 forward primer (labelled IRD 700), and 1.5 pmol M13 reverse primer (labelled IRD 800) in a total volume of 21 µL. A mix containing a combination of 4.5 µL of this master mix and 1.5 µl A, C, G or T reagent from the Thermosequenase Sequencing Kit (Amersham Pharmacia Biotech) is thermocycled using a Techne Genius (Fisher). The thermocycling conditions are as follows: an initial denaturation of 95°C for 30 sec followed by 20 cycles of [95°C for 10 sec, 54°C for 30 sec, 65°C for 30 sec] and 15 cycles of [95°C for 10 sec, 70°C for 30 sec]. Reactions are stopped by the addition of 4 µL stop solution (95% formamide, 10 mM EDTA, 0.05% pararosanaline). Samples are then denatured at 95°C for at least 3 min and 0.5 µL loaded onto a 41 cm sequencing gel. The sequencing gel comprises 7.5 mL Rapid Gel XL (Amersham Pharmacia Biotech), 21 g urea, 5 mL 10 X long run TBE and 28 mL H_2O. The solution is degassed for 15 min before adding 500 µL DMSO, 50 µL TEMED and 350 µL freshly made 10% ammonium persulphate. Gels are left to set for at least 2 hours before running. Running conditions are automatically specified for 41 cm gels using Base Image IR data collection software (1500 V, 35 mA, 40 W, 45°C, 25 frames, scan speed 3, prerun for 30 minutes in 1 X long run TBE)

Example 7

Huffman coding

Inefficiencies implicit in the use of substitution alphabets can be overcome to some extent by the use of Huffman coding.

Huffman encoding is an algorithm used to produce size-efficient codes to express data for which one has some kind of statistics concerning the occurrence of the symbols.

The operation of the algorithm is best explained by an example:

The letters of the alphabet {'a','b','c','d','e','f','g'} have known statistics for the frequency of occurrence for example {'a'=1,'b'=2,'c'=3,'d'=2,'e'=4,'f'=1,'g'=2} (ie 'e' occurs 4 times as often as 'a' or 'f'). We then write these two entities in columns next to each other sorted by frequency.

'e'	4
'c'	3
'b'	2
'd'	2
'g'	2
'a'	1
'f'	1

The algorithm commences by combining the entries in the table in pairs from the bottom. The frequencies of the pair under consideration are added and the two entries in the pair are replaced with an entry containing both the symbols and having a frequency equal to the sum of the two previous frequencies. Then the table is resorted (the order of the entries with equal frequencies is of no importance).

'e'	4
'c'	3
'a'+'f'	2
'd'	2
'g'	2
'b'	2

The process is repeated as follows until there is only one entry in the table:

1.

'e'	4
'c'	3
'b'	2
'd'	2
'g'	2
'a'	1
'f'	1

2.

'e'	4
'c'	3
'a'+'f'	2
'd'	2
'g'	2
'b'	2

3.

'g'+'b'	4
'e'	4
'c'	3
'a'+'f'	2
'd'	2

4.

'a'+'f'+'d'	4
'g'+'b'	4
'e'	4
'c'	3

5.

'e'+'c'	7
'a'+'f'+'d'	4
'g'+'b'	4

6.

'a'+'f'+'d'+'g'+'b'	8
'e'+'c'	7

7.

'a'+'f'+'d'+'g'+'b'+'e'+'c'	15

In order to extract the resulting codes from these tables, a "tree" is constructed by drawing lines between the cells that were merged into new cells. The root (the 15-cell here) splits into two branches. The top branch ('a','f','d','g','b') is assigned the value 0 and the bottom branch ('e','c') the value 1. This results in the codes for all symbols included in the top branch/cell starting with a 0 and the codes for the symbols in the bottom branch/cell starting with 1.

The construction of the tree is continued in this way. The top cell of (6) is split into the two bottom cells of (5), the top branch is assigned 0 and the bottom one 1. This results in symbols being incrementally assigned a number (0 or 1) depending on the route through the tree. At this state the code is as follows:

'a'	00
'b'	01
'c'	1
'd'	00
'e'	1
'f'	00
'g'	01

This process is repeated for each cell containing more than one symbol (ie 5) until there are no more tables. At this point the code in the example will be as follows

'a'	0000
'b'	01
'c'	11
'd'	001
'e'	10
'f'	0001
'g'	010

As can seen from the code above the most frequent letters are assigned the shortest codes.

Expanding on this example to demonstrate how the code is read consider the following symbol allocations

e: 000 a: 0010 d: 0011 r: 010 v: 01100 y: 01101 i: 0111
s: 1000 n: 1001 b: 10100 g: 10101 h: 1011 t: 111 space: 110.

In order to send the message "tra" the following symbols would be sent: 1110100010. The reader of the message would then start at the root of the tree and follow the numbered branches from the message until a letter is reached. This letter is recorded and the reader starts again at the root. This can be seen by reference to the figure 3.

Claims

1. A method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide wherein the embedded information is encoded, and decodable, by a mechanism which includes the use of Huffman coding and/or error checking.

2. A method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide, and decoding the embedded information by a mechanism that involves the use of Huffman coding and/or error checking.

3. A method according to claim 1 or 2, wherein the error checking mechanism comprises the use of a parity check or convolutional code.

4. A method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide wherein the embedded information corresponds to a sequence of digits in a quaternary code, the digits or combinations of digits having a predetermined meaning.

5. A method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide and treating it in such a way as to retrieve the embedded information as a series of digits in quaternary code, the digits or combinations of digits having a predetermined meaning.

6. A method of embedding information in a polynucleotide, the method comprising the step of forming a polynucleotide which comprises a plurality of information encoding units, each unit being cleavable from at least one other unit to generate a plurality of fragments, the absolute or relative lengths of which fragments denote the embedded information according to a predetermined relationship.

7. A method according to claim 6, wherein an information encoding unit is cleavable from at least one other unit by the action of a restriction endonuclease.

8. A method of decoding information embedded in a polynucleotide, the method comprising the step of isolating and/or optionally amplifying the polynucleotide, subjecting at least that part of the polynucleotide carrying the embedded information to digestion by at least one restriction endonuclease so as to generate a plurality of restriction fragments, wherein the absolute or relative lengths of the fragments have a predetermined meaning.

9. A method according to claim 8, wherein at least that part of the polynucleotide carrying the embedded information is subjected to digestion by at least two restriction endonucleases in a single reaction or in two separate reactions.

10. A method according to any one of the preceding claims wherein the polynucleotide comprises DNA.

11. A method according to any one of claims 6-10, wherein the polynucleotide contains a plurality of recognition/cleavage sites for at least two different restriction endonucleases.

12. A method according to any one of claims 6-11, wherein the polynucleotide comprises overlapping recognition/cleavage sites for restriction endonucleases.

13. A method according to any of claims 6-12, wherein the embedded information is decoded at least in part by determining the absolute and/or relative size of the fragments by PAGE.

14. A method according to any one of claims 6-13, wherein the absolute or relative size of the fragments denotes a series of digits in binary or quaternary code, digits or combinations of digits having a predetermined meaning.

15. A method according to any one of claims 1, 4 or 6, wherein the polynucleotide is introduced into a living host cell, preferably a plant cell, and preferably stably integrated into the genome thereof so as to be transmissible to progeny of the cell.

16. A method of embedding information in a polynucleotide as hereinbefore described and with reference to the accompanying drawings.

17. A method of decoding information embedded in a polynucleotide as hereinbefore described and with reference to the accompanying drawings.

Application No:	GB 0203140.9	**Examiner:**	Dr Patrick Purcell
Claims searched:	1-17	**Date of search:**	17 October 2002

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.T):

Int Cl (Ed.7):

Other: ONLINE: EPODOC, WPI, JAPIO, BIOSIS, MEDLINE

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
X	WO 01/00816 A1 (COMPLETE GENOMICS AS) see whole document, esp. page 30, lines 32-35, page 31, line 34-page 36, line 37, page 41, line 21-page 43, line 32	6-11, 13-15
A	WO 00/68431 A2 (MOUNT SINAI SCHOOL OF MEDICINE OF NEW YORK UNIVERSITY)	
X	WO 00/59917 A2 (RAUTHE) see WPI Abstract No: 2001-015649	6-11, 13-15
X	J. theor. Biol., Vol. 188, 1997, AJ Doig, "Improving the efficiency of the genetic code by varying the codon length- the perfect genetic code", 355-360	1, 2, 4, 5

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
		E	Patent document published on or after, but with priority date earlier than, the filing date of this application.
&	Member of the same patent family		